Frequency of frequencies distributions and size dependent exchangeable random partitions

Original Articles

Frequency of Frequencies Distributions and Size Dependent Exchangeable Random Partitions

Mingyuan Zhou, Stefano Favaro & Stephen G Walker

Page 0 | Received 24 Sep 2015, Accepted 29 Jul 2016, Accepted author version posted online: 26 Aug 2016

Download citation http://dx.doi.org/10.1080/01621459.2016.1222290 Crossmark

Abstract

Motivated by the fundamental problem of modeling the frequency of frequencies (FoF) distribution, this paper introduces the concept of a cluster structure to define a probability function that governs the joint distribution of a random count and its exchangeable random partitions. A cluster structure,

Frequency of Frequencies Distributions and Size Dependent Exchangeable Random Partitions

Mingyuan Zhou[†], Stefano Favaro[*], and Stephen G Walker[†]

mingyuan.zhou@mccombs.utexas.edu, stefano.favaro@unito.it, s.g.walker@math.utexas.edu

[†]The University of Texas at Austin, Austin, TX 78712, USA
[*]University of Torino and Collegio Carlo Alberto, 10134 Torino, Italy

Abstract

Motivated by the fundamental problem of modeling the frequency of frequencies (FoF) distribution, this paper introduces the concept of a cluster structure to define a probability function that governs the joint distribution of a random count and its exchangeable random partitions. A cluster structure, naturally arising from a completely random measure mixed Poisson process, allows the probability distribution of the random partitions of a subset of a population to be dependent on the population size, a distinct and motivated feature that makes it more flexible than a partition structure. This allows it to model an entire FoF distribution whose structural properties change as the population size varies. A FoF vector can be simulated by drawing an infinite number of Poisson random variables, or by a stick-breaking construction with a finite random number of steps. A generalized negative binomial process model is proposed to generate a cluster structure, where in the prior the number of clusters is finite and Poisson distributed, and the cluster sizes follow a truncated negative binomial distribution. We propose a simple Gibbs sampling algorithm to extrapolate the FoF vector of a population given the FoF vector of a sample taken without replacement from the population. We illustrate our results and demonstrate the advantages of the proposed models through the analysis of real text, genomic, and survey data.

Keywords: completely random measures, exchangeable cluster/partition probability functions, generalized negative binomial process, generalized Chinese restaurant sampling formula, species sampling.

1 Introduction

Characterizing a finite population whose individuals are partitioned into different classes is a fundamental research topic in physical, biological, environmental, and social sciences. One common problem is to estimate certain quantities of a sample taken from the population. For example, to disseminate survey data to the public, the government statistical agency has the responsibility to assess the risk for the disclosed microdata records to be matched to specific individuals of the surveyed population, based on the size and resolution of the microdata, while making them informative enough to be useful for education, research, business, and social welfare (Bethlehem et al., 1990; Fienberg and Makov, 1998; Manrique-Vallier and Reiter, 2012; Skinner and Elliot, 2002; Skinner and Shlomo, 2008).

In practice, one may not observe the population but only a sample taken from it. This brings another problem often more challenging to solve: to predict how the n individuals of a finite population are partitioned into different classes, on observing the partitions of a sample of $m < n$ individuals randomly taken from this population. For example, in high-throughput sequencing, one is often interested in estimating how many more new genomic sequences not found in the current sample would be detected if the sequencing depth is increased (Liu et al., 2014; Sims et al., 2014; Wang et al., 2009). To address this problem, one may define an appropriate procedure to extrapolate the random partitions of the population from the sample. One may also consider constructing a statistical model to fit the random partitions of the observed sample, with the assumption that the same model parameters inferred from the sample also apply to the population. The size-independent assumption, however, could considerably limit the flexibility of the selected statistical model. In addition, it could be restrictive to assume that the individuals of a random sample taken without replacement from a finite subpopulation are partitioned in the same way as those of a random sample taken without replacement from a larger population to which the subpopulation belongs.

To address all these problems under a coherent statistical framework, we will construct nonparametric Bayesian models to describe both the exchangeable random partitions of the population and those of a random sample taken without replacement from the population. The distribution of the random partitions of a sample will be constructed to be dependent on the population size, which is motivated by our observation that given the model parameters, the structural property of a sample's random partitions could strongly depend on both the size of the sample and that of the population.

The layout of the paper is as follows: In Section 1.1 we provide some background information. In Section 2, we discuss frequency of frequencies (FoF) distributions and introduce the new model

for constructing size dependent species sampling models. In Section 3 we apply the theory in Section 2 to the generalized negative binomial process and provide the asymptotics on both the number and sizes of clusters. We present real data applications in Section 4. We conclude the paper in Section 5 and provide the proofs in Appendix E.

1.1 Notation and preliminaries

Frequency of frequencies. Consider a finite population with n individuals from K different classes, and let $z_i \in \{1, \ldots, K\}$ denote the class individual i is assigned to, let $n_k = \sum_{i=1}^{n} \delta(z_i = k)$ denote the number of individuals in class k, and let $m_i = \sum_{k=1}^{K} \delta(n_k = i)$ denote the number of classes having i individuals in this finite population, where $\delta(x) = 1$ if the condition x is satisfied and $\delta(x) = 0$ otherwise. Thus, by definition, we have

$$K = \sum_{i=1}^{\infty} m_i, \quad n = \sum_{i=1}^{\infty} i m_i$$

almost surely (a.s.), and since $m_i = 0$ a.s. for all $i \geq n + 1$, it is also common to use $\sum_{i=1}^{n}$ to replace the infinite sum $\sum_{i=1}^{\infty}$ in the above equation. For example, we may represent $(z_1, \ldots, z_{14}) = (1, 2, 3, 4, 5, 5, 6, 6, 6, 6, 7, 7, 7, 7)$ as $(n_1, \ldots, n_7) = (1, 1, 1, 1, 2, 4, 4)$, or $\{m_1, m_2, m_4\} = \{4, 1, 2\}$ and $m_i = 0$ for $i \notin \{1, 2, 4\}$. Since m_i represents the frequency of the classes appearing i times, we refer the count vector $\mathcal{M} = \{m_i\}_i$ as the frequency of frequencies (FoF) vector, the distribution of which is commonly referred to as the FoF distribution (Good, 1953).

Exchangeable partition probability functions. Assuming the population size n is given, one may define a probability distribution to partition the n individuals into exchangeable random partitions, and hence generate a FoF vector by defining each partition as a class. Let $[m] := \{1, \ldots, m\}$ denote a subset of the set $[n] := \{1, \ldots, n\}$, where $m \leq n$. For a random partition $\Pi_m = \{A_1, \ldots, A_l\}$ of the set $[m]$, where there are l clusters and each individual $i \in [m]$ belongs to one and only one set A_k from Π_m, we denote $P(\Pi_m \,|\, n)$ as the marginal partition probability for $[m]$ when it is known the population size is n. Note that $P(\Pi_m \,|\, n) = P(z_1, \ldots, z_m \,|\, n)$ if individual i belongs to A_{z_i}.

If $P(\Pi_m \,|\, n)$ depends only on the number and sizes of the (A_k), regardless of their order, and the population size n, then it is referred to in this paper as a size dependent exchangeable partition probability function (EPPF) of Π_m. If $P(\Pi_m \,|\, m) = P(\Pi_m \,|\, n)$ for all $n \geq m$, then it is referred to as a size independent EPPF. Typical examples of size independent EPPFs include the Ewens sampling formula (Antoniak, 1974; Ewens, 1972), Pitman-Yor process (Perman et al., 1992; Pitman and Yor, 1997), and those governed by normalized random measures with independent increments (NRMIs) (Lijoi and Prünster, 2010; Regazzini et al., 2003). We provide a review on size independent EPPFs

in Appendix C. See Pitman (2006) for a detailed treatment of EPPFs.

Completely random measures. Let us denote G as a completely random measure (Kingman, 1967, 1993) defined on the product space $\mathbb{R}_+ \times \Omega$, where $\mathbb{R}_+ = \{x : x > 0\}$ and Ω is a complete separable metric space. It assigns independent infinitely divisible random variables $G(A_j)$ to disjoint Borel sets $A_j \subset \Omega$, with Laplace transforms

$$\mathbb{E}\left[e^{-\phi G(A)}\right] = \exp\left\{ - \int_{\mathbb{R}_+ \times A} (1 - e^{-\phi r}) \nu(dr d\omega) \right\}, \tag{1}$$

where $\nu(dr d\omega)$ is the Lévy measure. A random draw from G can be expressed as

$$G = \sum_{k=1}^{K} r_k \delta_{\omega_k}, \; K \sim \text{Poisson}(\nu^+), \; (r_k, \omega_k) \overset{iid}{\sim} \pi(dr d\omega),$$

where r_k is the weight of atom ω_k, $\nu^+ = \nu(\mathbb{R}_+ \times \Omega)$, and $\nu(dr d\omega) = \nu^+ \pi(dr d\omega)$. The completely random measure G is well defined if $\int_{\mathbb{R}_+ \times \Omega} \min\{1, r\} \nu(dr d\omega) < \infty$, even if the Poisson intensity ν^+ is infinite. In this paper, we consider homogenous completely random measures where the Lévy measure can be written as $\nu(dr d\omega) = \rho(dr) G_0(d\omega)$, where G_0 is a finite and continuous base measure over Ω.

The generalized gamma process $G \sim \text{gΓP}(G_0, a, 1/c)$ of Brix (1999), where $a < 1$ is a discount parameter and $1/c$ is a scale parameter, is defined with the Lévy measure as

$$\nu(dr d\omega) = \rho(dr) G_0(d\omega) = \frac{1}{\Gamma(1-a)} r^{-a-1} e^{-cr} \, dr \, G_0(d\omega). \tag{2}$$

A detailed description on the generalized gamma process is provided in Appendix D.

2 Bayesian modeling of frequency of frequencies

2.1 Frequency of frequencies distributions

The need to model the distributions of the class sizes $\{n_k\}_k$, or the FoF vector, arises in a wide variety of settings. For example, in computational linguistics and natural language processing, if we let n_k denote the frequency of the kth most frequent word in a text corpus, then $\ln(n_k)$ and $\ln(k)$ would be approximately linearly related according to Zipf's law (Zipf, 1949). Alternatively, if we let m_i denote the frequency of the words appearing i times, then $\ln(m_i)$ often appears to follow a straight line as a function of $\ln(i)$, as shown in Figures 1(a)-(d) for the words of four different novels. For many other natural and artificial phenomena, the FoF distributions also exhibit similar behavior in their tails, such as those on the number of citations of scientific papers, the degrees

of proteins in a protein-interaction network, and the peak gamma-ray intensity of solar flares, to name a few; see Newman (2005) and Clauset et al. (2009) for reviews. In addition, we find that the tails of the FoF distributions for the genomic sequences in high-throughput sequencing data and the classes of the microdata also often exhibit similar behaviors. For example, in Figure 1 are the FoF vectors for the words of four different novels[1], the RNA sequences of three different RNA-seq samples[2] provided by Frazee et al. (2011), and the classes of a microdata consists of 87,959 household records, shown in Table A.6 of Greenberg and Voshell (1990).

To illustrate how the characteristics of the FoF vector of a sample are related to the size of the sample, we show in Figure 2(a) the FoF distribution for all the words in the novel "The Adventures of Tom Sawyer" by Mark Twain on the logarithmic scale, and also plot the FoF distributions for 1/4, 1/16, 1/64, and 1/256 of the words taken without replacement from the novel, in Figures 2(b)-(e), respectively. We further show in Figure 3(a) the box plots of the slopes of the least squares regression lines fitted to the tails of these FoF vectors, and show in Figure 3(b) the box plots of the ratios of unit-size clusters (clusters of size one). In addition, we provide Figures A.1-A.2 in Appendix A as the analogous plots to Figures 2-3 for the FoF vectors for a high-throughput sequencing sample for the human transcriptome from a B cell line, as studied in Sultan et al. (2008). Note that to estimate the lower cutoff point and slope of the regression line, we use the software provided for Clauset et al. (2009), as described in detail in Appendix B.

It is clear from Figures 2-3 and A.1-A.2 that the slope of the fitted straight line and the ratio of unit-size clusters tend to decrease and increase, respectively, as the subsampling ratio decreases. Therefore, for a sample taken without replacement from a population, its estimated scaling parameter often clearly depends on the sample size. Moreover, it seems that a FoF distribution in some case could be more accurately described with a decreasing concave curve than with a straight line, such as those for the RNA sequences shown in Figures 1(e)-(g) and Figure A.1 in Appendix A. All these empirical observations motivate us to model the FoF distribution with a statistical model that could model the entire FoF distribution of a finite population, and more importantly, could take both the population and sample sizes into consideration, providing a principled way to extrapolate the FoF vector of a finite population given a random sample taken without replacement from the population.

[1] https://www.gutenberg.org/ebooks/

[2] http://bowtie-bio.sourceforge.net/recount/

2.2 Structure of the model

As discussed in Section 2.1 and shown in Figures 2-3 and A.1-A.2 in Appendix A, the structural property of a FoF distribution can strongly depend on n. Hence to use the same set of model parameters $\boldsymbol{\theta}$ to describe the FoF distributions for various sample sizes, we intend to construct a model that describe the distribution $P(\Pi_m \,|\, n, \boldsymbol{\theta})$, meaning that the EPPF and hence the FoF distribution for a sample of size m, taken without replacement from a population of size n, depends not only on the model parameters $\boldsymbol{\theta}$, but also on the population size n. To develop this theme, and to allow the mathematics to proceed in a neat way, and without forcing any restrictions, we first make n a random object within the model.

Here we describe how the random allocations of individuals to classes are distributed based on the independent random jumps of a completely random measure. With a random draw from a completely random measure expressed as $G = \sum_{k=1}^{K} r_k \delta_{\omega_k}$, by introducing a categorical latent variable z with $P(z = k \,|\, G) = r_k / G(\Omega)$, when a population of size n is observed we have

$$p(\boldsymbol{z} \,|\, G, n) = \prod_{i=1}^{n} \frac{r_{z_i}}{\sum_{k=1}^{K} r_k} = \left(\sum_{k=1}^{K} r_k \right)^{-n} \prod_{k=1}^{K} r_k^{n_k}, \tag{3}$$

where $\boldsymbol{z} = (z_1, \ldots, z_n)$ is a sequence of categorical random variables indicating the class memberships, $n_k = \sum_{i=1}^{n} \delta(z_i = k)$ is the number of data points assigned to category k, and $n = \sum_{k=1}^{K} n_k$. A random partition Π_n of $[n]$ is defined by the ties between the (z_i). So at this point, (3) is standard. Now (3) exhibits a lack of identifiabilty in that the scale of the (r_k) is arbitrary; the model is the same if we set $\widetilde{r}_k = \kappa r_k$ for any $\kappa > 0$. Hence, the total mass $\sum_{k=1}^{K} r_k$ is unidentified. Additionally, for the standard models, when G is integrated out, n disappears and we have $p(\boldsymbol{z})$ depending solely on the model parameters $\boldsymbol{\theta}$.

We solve both these issues by linking the population size n to the total random mass of G with a Poisson distribution, allowing n to depend on G via

$$p(n \,|\, G) = \text{Poisson}[G(\Omega)]. \tag{4}$$

Since the n data points are clustered according to the normalized random probability measure $G/G(\Omega)$, we have the equivalent sampling mechanism given by

$$p(n_k \,|\, G) = \text{Poisson}(r_k) \quad \text{independently for} \quad k = 1, 2, \ldots,$$

and, since $n = \sum_k n_k$, we obviously recover (4). We note here then that the prior model is for $p(n, G)$ and, consequently, $p(G \,|\, n)$ means G depends on n; *i.e.*, for each n we will have a different

random measure for G.

Therefore, we link directly the cluster sizes (n_k) to the weights (r_k) with independent Poisson distributions, which is in itself an appealing intuitive feature. The mechanism to generate a sample of arbitrary size is now well defined and G is no longer scaled freely. The new construction also allows $G(\Omega) = 0$, for which $n = 0$ a.s. Allowing $G(\Omega) = 0$ with a nonzero probability relaxes the requirement of $\nu^+ = \infty$ (*i.e.*, $K = \infty$ a.s.), a necessary condition to normalize a completely random measure (Lijoi and Prünster, 2010; Regazzini et al., 2003). For us we will not necessarily be assuming that $K = \infty$ a.s. In fact our model is such that $K = 0 \iff n = 0$, which is coherent, and, moreover, $P(K = 0 \,|\, n > 0) = 0$.

With G marginalized out from the G mixed Poisson process, the joint distribution of n and its exchangeable random partition Π_n is called an exchangeable cluster probability function (ECPF), which further leads to a FoF distribution that is shown to be an infinite product of Poisson distributions. On observing a population of size n, we are interested in the EPPF $P(\Pi_n \,|\, n, \boldsymbol{\theta})$ and, marginalizing over $n - m$ elements, we would consider $P(\Pi_m \,|\, n, \boldsymbol{\theta})$. Note that distinct from a partition structure of Kingman (1978a,b) that requires $P(\Pi_m \,|\, n, \boldsymbol{\theta}) = P(\Pi_m \,|\, m, \boldsymbol{\theta})$ for all $n > m$, we no longer have or require this condition for exchangeable random partitions generated under a G mixed Poisson process, which will be referred to as a cluster structure.

We provide in Section 2.3 the general form for both $p(\boldsymbol{z}, n) = P(\Pi_n, n \,|\, \boldsymbol{\theta})$ and $p(\boldsymbol{z} \,|\, n) = P(\Pi_n \,|\, n, \boldsymbol{\theta})$, and make connections to previous work in Section 2.4 by letting G be drawn from the gamma process. We provide in Section 3 the specific case when G is drawn from the generalized gamma process $G \sim \text{g}\Gamma\text{P}(G_0, a, 1/c)$ and the asymptotics on the number and sizes of clusters as $n \to \infty$. In Section 4 we use MCMC methods to extrapolate the FoF vector of the population from a random sample taken without replacement from it.

2.3 Properties of the model

A key insight of this paper is that a completely random measure mixed Poisson process produces a cluster structure that is identical in distribution to (*i*) the one produced by assigning the total random count of the Poisson process into exchangeable random partitions, using the random probability measure normalized from that completely random measure, (*ii*) the one produced by assigning the total (marginal) random count n of the mixed Poisson process into exchangeable random partitions using an EPPF of Π_n, and (*iii*) the one produced by constructing a FoF vector, the ith element of which is generated from a Poisson distribution parameterized by a specific function of i. For example, when the generalized gamma process $G \sim \text{g}\Gamma\text{P}[G_0, a, p/(1-p)]$ is used as the completely random measure in this setting, our key discoveries are summarized in Figure 4, which will be

discussed further in Section 3.

In Theorem 1, we establish the marginal model for the (n_k) with G marginalized out. We provide the Lévy measure, ECPF, EPPF, FoF distribution, stick-breaking construction, and prediction rule in Corollaries 2-5. The proofs are provided in Appendix E.

Theorem 1 (Compound Poisson Process)**.** *It is that the G mixed Poisson process is also a compound Poisson process; a random draw of which can be expressed as*

$$X(\cdot) = \sum_{k=1}^{l} n_k \, \delta_{\omega_k}(\cdot) \quad \textit{with} \;\; l \sim \text{Poisson}\left[G_0(\Omega) \int_0^\infty (1 - e^{-r})\rho(dr)\right],$$

and independently

$$P(n_k = j) = \frac{\int_0^\infty r^j e^{-r} \rho(dr)}{j! \int_0^\infty (1 - e^{-r})\rho(dr)} \;\; \textit{for} \;\; j = 1, 2, \ldots$$

where $\int_0^\infty (1 - e^{-r})\rho(dr) < \infty$ *is a condition required for the characteristic functions of G to be well defined,* $\omega_k \overset{iid}{\sim} g_0$*, and* $g_0(d\omega) = G_0(d\omega)/G_0(\Omega)$*.*

Corollary 2. *The Lévy measure of the G mixed Poisson process can be expressed as*

$$\nu(dnd\omega) = \sum_{j=1}^{\infty} \int_0^\infty \frac{r^j e^{-r}}{j!} \rho(dr) \, \delta_j(dn) G_0(d\omega).$$

The compound Poisson representation dictates the model to have a Poisson distributed finite number of clusters, whose sizes follow a positive discrete distribution. The mass parameter $\gamma_0 = G_0(\Omega)$ has a linear relationship with the expected number of clusters, but has no direct impact on the cluster-size distribution in the prior. Note that a draw from G contains $K < \infty$ or $K = \infty$ atoms a.s., but only l of them would be associated with nonzero counts if G is mixed with a Poisson process. Since the cluster indices are unordered and exchangeable, without loss of generality, in the following discussion, we relabel the atoms with nonzero counts in order of appearance from 1 to l and then $z_i \in \{1, \ldots, l\}$ for $i = 1, \ldots, n$, with $n_k > 0$ if and only if $1 \le k \le l$ and $n_k = 0$ if $k > l$.

Corollary 3 (Exchangeable Cluster/Partition Probability Functions)**.** *The model has a fully factorized exchangeable cluster probability function (ECPF) as*

$$p(\boldsymbol{z}, n \,|\, \gamma_0, \rho) = \frac{\gamma_0^l}{n!} \exp\left\{\gamma_0 \int_0^\infty (e^{-r} - 1)\rho(dr)\right\} \prod_{k=1}^{l} \int_0^\infty r^{n_k} e^{-r} \rho(dr),$$

the marginal distribution for the population size $n = X(\Omega)$ has probability generating function

$$\mathbb{E}[t^n \mid \gamma_0, \rho] = \exp\left\{\gamma_0 \int_0^\infty (e^{-(1-t)r} - 1)\rho(dr)\right\}$$

and probability mass function $p_N(n \mid \gamma_0, \rho) = \frac{d^n(\mathbb{E}[t^n \mid \gamma_0, \rho])}{n! dt^n}\Big|_{t=0}$, *and an exchangeable partition probability function (EPPF) of* Π_n *as*

$$p(\boldsymbol{z} \mid n, \gamma_0, \rho) = p(\boldsymbol{z}, n \mid \gamma_0, \rho)/p_N(n \mid \gamma_0, \rho).$$

The proof of this is straightforward given the representation in Theorem 1 and given the one-to-many-mapping combinatorial coefficient taking $(n_1, \ldots, n_l, l)$ to $(z_1, \ldots, z_n, n)$ is

$$\frac{l!}{n!} \prod_{k=1}^{l} n_k! \,.$$

Corollary 4 (Frequency of Frequencies Distribution)**.** *Let $\mathcal{M} = \{m_i\}_i$ be the frequency of frequencies (FoF) vector, where $m_i = \sum_{k=1}^{l} \delta(n_k = i)$ is the number of distinct types of size i, $\sum_{i=1}^{\infty} m_i = l$, and $\sum_{i=1}^{\infty} i m_i = n$. For the G mixed Poisson process, we can generate a random sample of $\mathcal{M}$ by drawing each of its element independently as*

$$m_i \sim \text{Poisson}\left(m_i; \frac{\gamma_0 \int_0^\infty r^i e^{-r} \rho(dr)}{i!}\right) \tag{5}$$

for $i \in \{1, 2, \ldots\}$. Alternatively, we may first draw

$$l \sim \text{Poisson}\left(\gamma_0 \int_0^\infty (1 - e^{-r})\rho(dr)\right)$$

as the total number of distinct clusters (species) with nonzero counts, then draw m_i sequentially using a stick-breaking construction as

$$m_i \mid l, m_1, \ldots, m_{i-1} \sim \text{Binomial}\left(l - \sum_{t=1}^{i-1} m_t, \frac{\frac{\int_0^\infty r^i e^{-r} \rho(dr)}{i!}}{\sum_{t=i}^{\infty} \frac{\int_0^\infty r^t e^{-r} \rho(dr)}{t!}}\right) \tag{6}$$

for $i = 1, 2, \ldots$ until $l = \sum_{t=1}^{i} m_i$, and further let $m_{i+\kappa} = 0$ for all $\kappa \in \{1, 2, \ldots\}$.

Corollary 5 (Prediction Rule)**.** *Let l^{-i} represent the number of clusters in $\boldsymbol{z}^{-i} := \boldsymbol{z} \backslash z_i$ and $n_k^{-i} := \sum_{j \neq i} \delta(z_j = k)$. We can express the prediction rule of the model as*

$$P(z_i = k \,|\, \boldsymbol{z}^{-i}, n, \gamma_0, \rho) \propto \begin{cases} \dfrac{\int_0^\infty r^{n_k^{-i}+1} e^{-r} \rho(dr)}{\int_0^\infty r^{n_k^{-i}} e^{-r} \rho(dr)}, & \text{for } k = 1, \ldots, l^{-i}; \\ \gamma_0 \displaystyle\int_0^\infty r e^{-r} \rho(dr), & \text{if } k = l^{-i} + 1. \end{cases}$$

This prediction rule can be used to simulate an exchangeable random partition of $[n]$ *via Gibbs sampling*.

2.4 Related work

To make connections to previous work, let us first consider the special case that G is a gamma process with Lévy measure $\nu(drd\omega) = r^{-1}e^{-p^{-1}(1-p)r}drG_0(d\omega)$, which is a special case of the generalized gamma process $G \sim \text{g}\Gamma\text{P}[G_0, a, p/(1-p)]$ with $a = 0$. This G mixed Poisson process is defined as the negative binomial process $X \sim \text{NBP}(G_0, p)$ in Zhou and Carin (2015). For $X \sim \text{NBP}(G_0, p)$, with Corollary 2, the Lévy measure can be expressed as $\nu(dnd\omega) = \sum_{j=1}^\infty j^{-1}p^j\delta_j(dn)G_0(d\omega)$. With Corollary 3, we have the ECPF $p(\boldsymbol{z}, n \,|\, \gamma_0, p) = (n!)^{-1}p^n(1-p)^{\gamma_0}\gamma_0^l \prod_{k=1}^l \Gamma(n_k)$ and probability mass function (PMF) $p_N(n \,|\, \gamma_0, p) = \frac{\Gamma(n+\gamma_0)}{\Gamma(\gamma_0)}p^n(1-p)^{\gamma_0}$, which is the PMF of the negative binomial (NB) distribution $n \sim \text{NB}(\gamma_0, p)$. Thus the EPPF for X can be expressed as

$$p(\boldsymbol{z} \,|\, \gamma_0) = \frac{p(\boldsymbol{z}, n \,|\, \gamma_0, p)}{p_N(n \,|\, \gamma_0, p)} = \frac{\Gamma(\gamma_0)\gamma_0^l}{\Gamma(n+\gamma_0)} \prod_{k=1}^l \Gamma(n_k), \tag{7}$$

which is the EPPF of the Chinese restaurant process (CRP) (Aldous, 1983), a variant of the widely used Ewens sampling formula (Blackwell and MacQueen, 1973; Ewens, 1972).

For the CRP, multiplying its EPPF $p(\boldsymbol{z} \,|\, \gamma_0)$ by the PMF of $n \sim \text{NB}(\gamma_0, p)$ leads to the ECPF, and as in Corollary 4, further multiplying its ECPF with the combinatorial coefficient $n!/[\prod_{i=1}^n (i!)^{m_i} m_i!]$ leads to the distribution of a FoF vector $\mathcal{M} = \{m_i\}_i$ as

$$p(\mathcal{M}, n \,|\, \gamma_0, p) = \left\{ \prod_{i=1}^\infty \text{Poisson}\left(m_i; \gamma_0 \frac{p^i}{i}\right) \right\} \times \delta\left(n = \sum_{i=1}^\infty i m_i\right),$$

which can be generated by simulating countably infinite Poisson random variables, or using a stick-breaking construction that first draws $l \sim \text{Poisson}[-\gamma_0 \ln(1-p)]$ number of of nonempty clusters, and then draws m_i sequentially

$$m_i \,|\, l, m_1, \ldots, m_{i-1} \sim \text{Binomial}\left(l - \sum_{t=1}^{i-1} m_t, \frac{i^{-1}p^i}{-\ln(1-p) - \sum_{t=1}^{i-1} t^{-1}p^t}\right) \tag{8}$$

for $i = 1, 2, \ldots$ until $l = \sum_{t=1}^{i} m_i$, and further lets $m_{i+\kappa} = 0$ for all $\kappa \in \{1, 2, \ldots\}$.

The EPPF of the widely used Piman-Yor process (Pitman, 2006), with mass parameter γ_0 and discount parameter $a \in [0, 1)$, can be expressed as

$$P(\boldsymbol{z} \,|\, \gamma_0, a) = \frac{\Gamma(\gamma_0)}{\Gamma(n + \gamma_0)} \prod_{k=1}^{l} \frac{\Gamma(n_k - a)}{\Gamma(1 - a)} [\gamma_0 + (k - 1)a].$$

However, unless $a = 0$, it is unclear whether the Pitman-Yor process can be related to a FoF vector whose countably infinite elements simply follow the Poisson distributions. There exists the class of Gibbs-type EPPF that provides a generalization of the EPPF induced by the Pitman-Yor process. See Gnedin and Pitman (2006) for details and De Blasi et al. (2015) for a Bayesian nonparametric treatment.

Note that the ideas of mixing multiple group-specific Poisson processes with a gamma process, or mixing multiple group-specific negative binomial (NB) processes with a gamma or beta process have been exploited in Zhou and Carin (2015) to construct priors for mixed-membership modeling, and in Zhou et al. (2015) to construct priors for random count matrices. When the number of groups reduces to one, the NB process in Zhou and Carin (2015) and Zhou et al. (2015) becomes a special case of the generalized NB process to be thoroughly investigated in Section 3. Following the hierarchical construction in Zhou and Carin (2015) and Zhou et al. (2015), the proposed generalized NB process or other completely random measure mixed Poisson processes may also be extended to a multiple group setting to construct more sophisticated nonparametric Bayesian priors for both mixed-membership modeling and random count matrices.

Below we will study a particular process: the generalized NB process, whose ECPF and FoF distribution both have simple analytic expressions and whose exchangeable random partitions can not only be simulated via Gibbs sampling using the above prediction rule, but also be sequentially constructed using a recursively calculated prediction rule.

3 Generalized negative binomial process

In the following discussion, we study the generalized NB process (gNBP) model where $G \sim \mathrm{g\Gamma P}[G_0, a, p/(1-p)]$ with $a < 0$, $a = 0$, or $0 < a < 1$. Here we apply the results in Section 3 to this specific case. Using (2), we have $\int_0^\infty r^n e^{-r} \rho(dr) = \frac{\Gamma(n-a)}{\Gamma(1-a)} p^{n-a}$ and $\int_0^\infty (1 - e^{-r}) \rho(dr) = \frac{1-(1-p)^a}{a p^a}$. Marginalizing out $G(\Omega)$ from $n \,|\, \lambda \sim \mathrm{Poisson}[G(\Omega)]$ with $G \sim \mathrm{g\Gamma P}[\gamma_0, a, p/(1-p)]$, leads to a generalized NB distribution; *i.e.*, $n \sim \mathrm{gNB}(\gamma_0, a, p)$, with shape parameter γ_0, discount parameter $a < 1$, and probability parameter p. Denote by $\sum_*$ as the summation over all sets of positive integers $(n_1, \ldots, n_l)$ with $\sum_{k=1}^{l} n_k = n$. As derived in Appendix F, the PMF of the generalized NB

distribution can be expressed as

$$p_N(n \mid \gamma_0, a, p) = p^n e^{-\gamma_0 \frac{1-(1-p)^a}{ap^a}} \sum_{l=0}^{n} \gamma_0^l p^{-al} \frac{S_a(n,l)}{n!}, \tag{9}$$

where $S_a(n,l)$, as defined in detail in Appendix F, multiplied by a^{-l} are generalized Stirling numbers (Charalambides, 2005; Pitman, 2006).

Marginalizing out G in the generalized gamma process mixed Poisson process

$$X \mid G \sim \mathrm{PP}(G) \quad \text{and} \quad G \sim \mathrm{g\Gamma P}\left[G_0, a, p/(1-p)\right] \tag{10}$$

leads to a generalized NB process $X \sim \mathrm{gNBP}(G_0, a, p)$, such that for each $A \subset \Omega$, $X(A) \sim \mathrm{gNB}(G_0(A), a, p)$. This process is also a compound Poisson process as

$$X(\cdot) = \sum_{k=1}^{l} n_k \delta_{\omega_k}(\cdot),\ l \sim \mathrm{Poisson}\Big(\gamma_0 \frac{1-(1-p)^a}{ap^a}\Big),\ n_k \overset{iid}{\sim} \mathrm{TNB}(a,p),\ \omega_k \overset{iid}{\sim} g_0, \tag{11}$$

where $\mathrm{TNB}(a,p)$ denotes a truncated NB distribution, with PMF

$$p_U(u \mid a, p) = \frac{\Gamma(u-a)}{u!\Gamma(-a)} \frac{p^u (1-p)^{-a}}{1-(1-p)^{-a}},\ u = 1, 2, \ldots. \tag{12}$$

Note that $\lim_{a\to 0} \frac{1-(1-p)^a}{ap^a} = -\ln(1-p)$ and $\lim_{a\to 0} \mathrm{TNB}(a,p)$ becomes the logarithmic distribution with parameter p (Fisher et al., 1943; Johnson et al., 2005; Quenouille, 1949). The Lévy measure of the gNBP can be expressed as $\nu(dnd\omega) = \sum_{j=1}^{\infty} \frac{\Gamma(j-a)}{j!\Gamma(1-a)} p^{j-a} \delta_j(dn) G_0(d\omega)$.

The ECPF of the gNBP model is given by

$$p(\boldsymbol{z}, n \mid \gamma_0, a, p) = \frac{1}{n!} e^{-\gamma_0 \frac{1-(1-p)^a}{ap^a}} \gamma_0^l p^{n-al} \prod_{k=1}^{l} \frac{\Gamma(n_k - a)}{\Gamma(1-a)}, \tag{13}$$

which is fully factorized and will be used as the likelihood to infer γ_0, a, and p. The EPPF of Π_n is the ECPF in (13) divided by the marginal distribution of n in (9), given by

$$p(\boldsymbol{z} \mid n, \gamma_0, a, p) = \frac{\gamma_0^l p^{-al}}{\sum_{\ell=0}^{n} \gamma_0^\ell p^{-a\ell} S_a(n,\ell)} \prod_{k=1}^{l} \frac{\Gamma(n_k - a)}{\Gamma(1-a)}. \tag{14}$$

We define the EPPF in (14) as the generalized Chinese restaurant sampling formula (gCRSF), and we denote a random draw under this EPPF as

$$\boldsymbol{z} \mid n \sim \mathrm{gCRSF}(n, \gamma_0, a, p).$$

The conditional distribution of the number of clusters in a population of size n can be expressed as

$$p_L(l \mid n, \gamma_0, a, p) = \frac{1}{l!} \sum_{*} \frac{n!}{\prod_{k=1}^{l} n_k!} p(\boldsymbol{z} \mid n, \gamma_0, a, p) = \frac{\gamma_0^l p^{-al} S_a(n, l)}{\sum_{\ell=0}^{n} \gamma_0^\ell p^{-a\ell} S_a(n, \ell)}. \tag{15}$$

Recall that $m_i = \sum_{k=1}^{l} \delta(n_k = i)$ represents the number of distinct types of size i, with $\sum_{i=1}^{\infty} m_i = l$ and $\sum_{i=1}^{\infty} i m_i = n$. With Corollary 4, we can express the joint distribution of n and $\mathcal{M}$, under the constraint that $n = \sum_{i=1}^{\infty} i m_i$, as

$$p(\mathcal{M}, n \mid \gamma_0, a, p) = \left\{ \prod_{i=1}^{\infty} \text{Poisson}\left(m_i; \frac{\Gamma(i-a)\gamma_0 p^{i-a}}{\Gamma(1-a) i!} \right) \right\} \times \delta\left(n = \sum_{i=1}^{\infty} i m_i \right), \tag{16}$$

where we apply the fact that $\sum_{i=1}^{n} \frac{\Gamma(i-a)}{i!\Gamma(-a)} p^i (1-p)^{-a} = 1 - (1-p)^{-a}$ for $a < 1$. Thus to generate a cluster structure governed by the generalized negative binomial process, one may draw $m_i \sim \text{Poisson}\left(\frac{\Gamma(i-a)\gamma_0 p^{i-a}}{\Gamma(1-a) i!} \right)$ independently for each i, or first draw

$$l \sim \text{Poisson}\left(\gamma_0 \frac{1 - (1-p)^a}{a p^a} \right) \tag{17}$$

number of unique partitions (species), and then draw m_i for $i \geq 1$ using

$$m_i \mid l, m_1, \ldots, m_{i-1} \sim \text{Binomial}\left(l - \sum_{t=1}^{i-1} m_t, \frac{\frac{\Gamma(i-a) p^i}{i!}}{\sum_{t=i}^{\infty} \frac{\Gamma(t-a) p^t}{t!}} \right) \tag{18}$$

until $l = \sum_{t=1}^{i} m_t$. Note that in the prior, $\mathbb{E}[m_i] = \left(\frac{\Gamma(i-a)\gamma_0 p^{i-a}}{\Gamma(1-a) i!} \right)$ and hence, using the property of the gamma function, we have

$$\ln(\mathbb{E}[m_i]) \;\sim\; -(a+1)\ln(i) + \ln(p) i$$

as $i \to \infty$. Thus if $p \to 1$, we may consider $a + 1$ as a power-law scaling parameter.

Note that if $a \to 0$, we recover from (16) the logarithmic series of Fisher et al. (1943), as also discussed in Anscombe (1950) and Watterson (1974), and we recover from (14) the EPPF for the CRP, as shown in (7). When $a \neq 0$, we generalize CRP by making the EPPF be dependent on the population size n. This generalization differs from those in Ishwaran and James (2003) and Cerquetti (2008), where the EPPFs are independent of n.

The prediction rule for the EPPF in (14) can be expressed as

$$P(z_i = k \mid \boldsymbol{z}^{-i}, n, \gamma_0, a, p) \propto \begin{cases} n_k^{-i} - a, & \text{for } k = 1, \ldots, l^{-i}; \\ \gamma_0 p^{-a}, & \text{if } k = l^{-i} + 1. \end{cases} \tag{19}$$

This prediction rule can be used in a Gibbs sampler to simulate an exchangeable random partition $z \,|\, n \sim \text{gCRSF}(n, \gamma_0, a, p)$ of $[n]$. As it is often unclear how many Gibbs sampling iterations are required to generate an unbiased sample from this EPPF, below we present a sequential construction for this EPPF to directly generate an unbiased sample.

Marginalizing out z_n from (14), we have

$$\begin{aligned} p(z_{1:n-1} \,|\, n, \gamma_0, a, p) &= p(z_{1:n-1} \,|\, n-1, \gamma_0, a, p) \\ &\quad \times \frac{\sum_{\ell=0}^{n-1} \gamma_0^{\ell} p^{-a\ell} S_a(n-1, \ell)}{\sum_{\ell=0}^{n} \gamma_0^{\ell} p^{-a\ell} S_a(n, \ell)} \left[\gamma_0 p^{-a} + (n-1) - a l_{(n-1)}\right], \end{aligned}$$

where $z_{1:i} := \{z_1, \ldots, z_i\}$, $l_{(i)}$ denotes the number of partitions in $z_{1:i}$, and $l_{(n)} = l$. Further marginalizing out $z_{n-1}, \ldots, z_{i+1}$, we have

$$\begin{aligned} p(z_{1:i} \,|\, n, \gamma_0, a, p) &= p(z_{1:i} \,|\, i, \gamma_0, a, p) \frac{\sum_{\ell=0}^{i} \gamma_0^{\ell} p^{-a\ell} S_a(i, \ell)}{\sum_{\ell=0}^{n} \gamma_0^{\ell} p^{-a\ell} S_a(n, \ell)} R_{n,\gamma_0,a,p}(i, l_{(i)}) \\ &= \frac{R_{n,\gamma_0,a,p}(i, l_{(i)}) \gamma_0^{l_{(i)}} p^{-a l_{(i)}}}{\sum_{\ell=0}^{n} \gamma_0^{\ell} p^{-a\ell} S_a(n, \ell)} \prod_{k \,:\, n_{k,(i)} > 0} \frac{\Gamma(n_{k,(i)} - a)}{\Gamma(1-a)}, \end{aligned} \tag{20}$$

where $n_{k,(i)} := \sum_{j=1}^{i} \delta(z_j = k)$; $R_{n,\gamma_0,a,p}(i, j) = 1$ if $i = n$ and is recursively calculated for $i = n-1, n-2, \ldots, 1$ with

$$R_{n,\gamma_0,a,p}(i, j) = R_{n,\gamma_0,a,p}(i+1, j)(i - a j) + R_{n,\gamma_0,a,p}(i+1, j+1)\gamma_0 p^{-a}. \tag{21}$$

We name (20) as a size-dependent EPPF as its distribution on an exchangeable random partition of $[i]$ is a function of the population size n. Note that if $a = 0$, the EPPF becomes the same as that of the Chinese restaurant process and no longer depends on n.

In Appendix F, we show the sequential prediction rule of the generalized Chinese restaurant sampling formula that constructs Π_{i+1} from Π_i in a population of size n by assigning element $(i+1)$ to $A_{z_{i+1}}$, and show the predictive distribution of $z_{i+1:n}$ given $z_{1:i}$, the population size n, and model parameters.

In summary, a draw from the generalized NB process (gNBP) represents a cluster structure with a Poisson distributed finite number of clusters, whose sizes follow a truncated NB distribution. Marginally, the population size follows a generalized NB distribution. These three count distributions and the prediction rule are determined by a discount, a probability, and a mass parameter, which together with i are used to parameterize the Poisson rate for the random number of clusters of size i for the FoF distribution. These parameters are convenient to infer using the

fully factorized ECPF. Since $P(\Pi_m \,|\, n) = P(\Pi_m \,|\, m)$ is often not true for $n > m$, the EPPF of the gNBP, which is derived by applying Bayes' rule on the ECPF and the generalized NB distribution, generally violates the addition rule required in a partition structure and hence is dependent on the population size. This size dependent EPPF is referred to as the generalized Chinese restaurant sampling formula. To generate an exchangeable random partition of $[n]$ under this EPPF, we show we could use either a Gibbs sampler or a recursively-calculated sequential prediction rule.

We conclude this section by investigating the large n asymptotic behavior of both the number of clusters $p_L(l \,|\, n, \gamma_0, a, p)$ shown in (15) and the sizes of clusters $p(\mathcal{M} \,|\, n, \gamma_0, a, p) = p(\mathcal{M}, n \,|\, \gamma_0, a, p)/p_N(n \,|\, \gamma_0, a, p)$ which can be obtained with (16) and (9). An interesting question to answer is if we fix the model parameters γ_0, a, and p, where $0 < \gamma_0 < \infty$, $a < 1$, and $0 < p < 1$, and assume the population size n is given, how $l_{(n)}$, the cluster number, and $M_{i,n}$, the number of clusters of size i, would behave as the population size n approaches infinity. We summarize our findings in Table 1 and provide the details in Appendices G and H. Table 1 characterizes three asymptotic regimes according to the choice of the parameter a, that is $a \in (0, 1)$, $a = 0$, and $a \in \{-1, -2, \ldots\}$.

For $a = 0$ the distribution (15) coincides with the distribution of the number of clusters in a sample of size n from a Dirichlet process. Hence, the large n asymptotic behavior of $l_{(n)}$ is known from Korwar and Hollander (1973) whereas the large n asymptotic behavior of $M_{i,n}$ is known from Ewens (1972).

For any $a \in (0, 1)$ the number of clusters minus one, $l_{(n)}-1$, converges weakly to Poisson$[\gamma_0/(ap^a)]$, whereas $M_{i,n}$ converges weakly to Poisson$\left(\frac{\Gamma(i-a)\gamma_0 p^{-a}}{\Gamma(1-a)i!}\right)$. Note that, for any $a \in (0, 1)$, $a\frac{\Gamma(i-a)}{\Gamma(1-a)i!}$ is a proper probability distribution over the natural numbers, that is $a\frac{\Gamma(i-a)}{\Gamma(1-a)i!} \in (0, 1)$ for any $i \geq 1$ and $\sum_{i=1}^{\infty} a\frac{\Gamma(i-a)}{\Gamma(1-a)i!} = 1$. In other terms, for large n the number $M_{i,n}$ of clusters of size i becomes a proportion $a\frac{\Gamma(i-a)}{\Gamma(1-a)i!}$ of $l_{(n)} - 1$, and such a proportion decreases with the index i. It is also interesting to notice that the logarithmic of $\frac{\Gamma(i-a)\gamma_0 p^{-a}}{\Gamma(1-a)i!}$ can be approximated by

$$-(a + 1)\ln(i) + C$$

when i is large, where the coefficient $C = \ln\left(\frac{\gamma_0 p^{-a}}{\Gamma(1-a)}\right)$ is not related to the index i. Thus we may consider $a + 1$ as a power-law scaling parameter as $n \to \infty$.

Finally, for any $a \in \{-1, -2, \ldots\}$ the number of clusters rescaled by $n^{-a/(1-a)}$ converges weakly to the constant $\frac{(\gamma_0 p^{-a})^{\frac{1}{1-a}}}{-a}$, whereas $M_{i,n}$ converges weakly to Poisson$\left(\frac{\Gamma(i-a)\gamma_0 p^{-a}}{\Gamma(1-a)i!}\right)$. Note that, differently from the case $a \in (0, 1)$, for any $a \in \{-1, -2, \ldots\}$, $\sum_{i=1}^{\infty} a\frac{\Gamma(i-a)}{\Gamma(1-a)i!} = +\infty$, that is $a\frac{\Gamma(i-a)}{\Gamma(1-a)i!}$ is not a probability distribution over the natural numbers. In particular, $a\frac{\Gamma(i-a)}{\Gamma(1-a)i!}$ is a constant when $a = -1$ and increases with the index i when $a \in \{-2, -3, \ldots\}$.

4 Illustrations

Species abundance data of a population is usually represented with a FoF vector as $\mathcal{M} = \{m_i\}_i$, where m_i denotes the number of species that have been observed i times in the population. As discussed before, this data can also be converted into a sequence of cluster indices $\boldsymbol{z} = (z_1, \ldots, z_n)$ or a cluster-size vector $(n_1, \ldots, n_l)$, where n_k is the number of individuals in cluster k, $n = \sum_i i m_i = \sum_{k=1}^{l} n_k$ is the size of the population and $l = \sum_i m_i$ is the number of distinct clusters in the population. For example, we may represent $\{m_1, m_2, m_3\} = \{2, 1, 2\}$ as $\boldsymbol{z} = (1, 2, 3, 3, 4, 4, 4, 5, 5, 5)$ or $(n_1, \ldots, n_5) = (1, 1, 2, 3, 3)$. For species frequency counts, we use (13) as the likelihood for the model parameters $\boldsymbol{\theta} = \{\gamma_0, a, p\}$. With appropriate priors imposed on $\boldsymbol{\theta}$, we use MCMC to obtain posterior samples $\boldsymbol{\theta}^{(j)} = \{\gamma_0^{(j)}, a^{(j)}, p^{(j)}\}$. The details of MCMC update equations are provided in Appendix I.

To understand the structural properties of the population, one often has to make a choice between taking more but smaller size samples and taking fewer but larger size samples. For example, in high-throughput sequencing, to increase the number of detected sequences given a fixed budget, one may need to decide whether to reduce the sequencing depth per sample to allow collecting more biological replicates (Sims et al., 2014). These motivate us to consider the fundamental problem of extrapolating the FoF vector of a sample, taken without replacement from the population, to reconstruct the FoF vector of the population. This extrapolation problem is readily answered under our framework by $p(z_{i+1:n} \,|\, z_{1:i}, n, \gamma_0, a, p)$ in (F.8), which shows the joint distribution of the cluster indices of the unobserved $n - i$ individuals of the population given the observed clusters indices $(z_1, \ldots, z_i)$ of the sample of size i, the population size n, and the model parameters. To reconstruct $(z_{i+1}, \ldots, z_n)$, one can either use (19) to sequentially construct the vector from z_{i+1} to z_n, or randomly initialize the vector and then use (F.7) in a Gibbs sampling algorithm. For a population with tens of thousands or millions of individuals, we prefer the second method as it is often more computationally efficient.

We consider the novel "The Adventures of Tom Sawyer" by Mark Twain, with a total of $n = 77,514$ words from $l = 7,772$ terms; the novel "The Adventures of Sherlock Holmes" by Arthur Conan Doyle, with a total of $n = 106,007$ words from $l = 7,896$ terms; the high-throughput sequencing dataset studied in Sultan et al. (2008), with a total of $n = 418,650$ sequences from $l = 6,712$ unique sequences; the high-throughput sequencing dataset studied in Core et al. (2008), with a total of $n = 125,794$ sequences from $l = 7,124$ unique sequences; and the mircodata provided in Table A.6 of Greenberg and Voshell (1990), with a total of $n = 87,959$ household records from $l = 929$ groups. We randomly take $1/32$, $1/16$, $1/8$, $1/4$, or $1/2$ of the individuals without replacement from the population to form a sample $(z_1, \ldots, z_i)$, where i is the sample size, from

which we use Gibbs sampling to simulate the indices of the remaining individuals $(z_{i+1}, \ldots, z_n)$, where n is the population size. In each Gibbs sampling iteration, we draw $T = 5$ times the indices in $\{z_{i+1}, \ldots, z_n\}$ in a random order using (F.7) and then sample the model parameters γ_0, a, and p once.

For comparison, we consider using the software provide for Clauset et al. (2009) to estimate a lower cutoff point $i_{\min}$ and a scaling parameter α from a random sample taken without replacement from the finite population, and then find $-\alpha_h$, the slope of the least squares line fitting the first $i_{\min} - 1$ FoF points of the random sample on the log-log plot. We then fit a straight line to the population FoF points $\{\ln i, \ln(m_i)\}_{i<i_{\min}}$, with $-\alpha_h$ as the slope and $[\sum_{i\in I_h}(\ln(m_i) + \alpha_h \ln(m_i)]/|I_h|$ as the intercept, where $I_h = \{i : 1 \le i < i_{\min}, m_i >= 1\}$, and another straight line to the population FoF points $\{\ln i, \ln(m_i)\}_{i\ge i_{\min}}$, with $-\alpha$ as the slope and $[\sum_{i\in I_t}(\ln(m_i) + \alpha \ln(m_i)]/|I_t|$ as the intercept, where $I_t = \{i : i \ge i_{\min}, m_i >= 3\}$. We emphasize that this least squares (LS) procedure is merely used as a baseline, which refits the population FoF points under the assumption that $i_{\min}$, α_h, and α all all stay unchanged as the sample size varies; it may fit the tail well, but may perform poorly in fitting the center part of a FoF distribution.

We also make comparisons with the Pitman-Yor process (Perman et al., 1992; Pitman, 2006; Pitman and Yor, 1997), a widely used nonparametric Bayesian prior with a size independent EPPF that $P(\Pi_m \,|\, \gamma_0, a, m) = P(\Pi_m \,|\, \gamma_0, a, n)$ for all $n \ge m$, where γ_0 and a are the concentration and discount parameters, respectively, for the Pitman-Yor process. We describe a Gibbs sampling algorithm in Appendix I, using data augmentation techniques developed in Teh (2006). In addition, we also consider the Chinese restaurant process.

For all MCMC based algorithms, we consider 1000 iterations and collect the last 500 samples, for each of which we convert the cluster index vector $(z_1, \ldots, z_n)$ to a population FoF vector, and take the average of all the 500 collected vectors, denoted by $\widehat{\mathcal{M}} = (\hat{m}_1, \ldots, \hat{m}_n)$, as the posterior mean of the population FoF vector, given the sample $(z_1, \ldots, z_i)$ and the population size n. Using the observed population FoF vector $\mathcal{M}$, we measure the extrapolation performance using the root mean squared error (RMSE), defined as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{100} \delta(m_i > 0)\,[\ln(m_i) - \ln(\hat{m}_i)]^2}{\sum_{i=1}^{100} \delta(m_i > 0)}} \tag{22}$$

and the chi-squared test statistic, defined as

$$\chi^2 = \frac{(\sum_{i=50}^{n} m_i - \sum_{i=50}^{n} \hat{m}_i)^2}{\sum_{i=50}^{n} \hat{m}_i} + \sum_{i=1}^{49} \frac{(m_i - \hat{m}_i)^2}{\hat{m}_i}. \tag{23}$$

The RMSE and chi-squared test statistic measure the distances between the observed population FoF vector and the extrapolated FoF vector in the logarithmic and original scales, respectively. Examining the trace plots of the inferred model parameters, we find that 1000 MCMC iterations are sufficient for both the Pitman-Yor and generalized NB process, as the Markov chains appear to converge fast and mix well in all experiments. We provide example trace plots for three different datasets in Figures A.3-A.5 of Appendix A.

Shown in Figure 5 are the posterior means of the population FoF vectors extrapolated from sample FoF vectors for "The Adventures of Tom Sawyer" by Mark Twain, using least squares (LS) lines fitted to the population FoF points on the log-log plots, using the Pitman-Yor process, or using the generalized negative binomial process under various settings of the discount parameter a. Shown in Figure 6 are the corresponding RMSEs and chi-squared test statistics. Note that the slopes of these LS lines are estimated from the sample FoF vectors, whereas the intercepts are obtained by refitting these straight lines to the population FoF vectors. Thus the LS procedure is appropriate for fitting the data but impractical for out-of-sample prediction. The results of the Chinese restaurant process are almost identical to these of the generalized negative binomial process with $a = 0$, and hence are omitted from these figures. Figures 7-8 are analogous plots to Figures 5-6 for a high-throughput RNA-seq data studied in Sultan et al. (2008), and Figures 9-10 are analogous plots to Figures 5-6 for a microdata. In Appendix A, we also provide corresponding Figures A.6-A.7 for "The Adventures of Sherlock Holmes" by Arthur Conan Doyle, and Figures A.8-A.9 for a high-throughput RNA-seq data studied in Core et al. (2008).

As shown in Figures 5-10 and Figures A.6-A.9 of Appendix A, the LS refitting procedure, impractical for real applications, consistently underperforms both the Pitman-Yor process and the gNBP with $a < 1$, and may perform poorly if the population FoF vector appears to follow a decreasing concave curve. The gNBP with $a = -1$ appears to strongly discourage the frequencies of small-size clusters. Although it has poor performance for all the data considered in the paper, it shows that $a = -1$ or even smaller values could be used for certain applications that favor the population FoF vector to follow a concave shape. Both the gNBP with $a = 0$, with almost identical performance to that of the Chinese restaurant process, and the gNBP with $a < 0$ perform well on both RNA-seq genomic data, each of whose population FoF vectors clearly follows a decreasing concave curve, but clearly underperform both the Pitman-Yor process and gNBP with $a < 1$ on the other three datasets, whose population FoF vectors more closely follow decreasing straight lines. The Pitman-Yor process performs well for all datasets, but in general clearly underperforms the gNBP with $a < 1$. In addition to the five datasets, we have also examined the other three datasets shown in Figure 1. Our observations on all these datasets consistently suggest that choosing the

gNBP, with a vary freely within $(-\infty, 1)$, achieves the performance that is either the best or close to the best, which is hence recommended as the preferred choice, if there is no clear prior information on how the population FoF vector is distributed.

5 Conclusions

We propose an infinite product of Poisson density functions to model the entire frequency of frequencies (FoF) distribution of a population consisting of a random number of individuals, and propose a size dependent exchangeable random partition function to model the FoF distribution of a population whose number of individuals is given. We first present a general framework that uses a completely random measure mixed Poisson process to support a FoF distribution, and then focus on studying the generalized negative binomial process constructed by mixing the generalized gamma process with the Poisson process. Our asymptotic analysis shows how the generalized negative binomial process can adjust its discount parameter to model different tail behaviors for the FoF distributions. On observing a single sample taken without replacement from a population, we propose a simple Gibbs sampling algorithm to extrapolate the FoF vector of the population from the FoF vector of that sample. The performance of the algorithm is demonstrated in estimating FoF vectors for text corpora, high-throughput sequencing data, and microdata, where a population typically consists of tens of thousands or millions of individuals. Since various kinds of statistics commonly used to characterize the properties of a population can often be readily calculated given the population FoF vector, being able to accurately model the FoF distributions of big datasets brings new opportunities to advance the state-of-the-art of a wide array of real discrete data applications, such as making comparisons between different text corpora, finding a good compromise between the depth and coverage of high-throughput sequencing for genomic data, estimating entropy in a nonparametric Bayesian manner, and assessing disclosure risk for microdata.

Acknowledgements

The authors thank the Associate Editor and three anonymous referees, whose invaluable comments and suggestions have helped us to improve the paper substantially. M. Zhou thanks Lawrence Carin, Fernando A. Quintana, Peter Müller for their comments on an earlier draft of this paper, and thanks Xiaoning Qian and Siamak Zamani Dadaneh for discussions on high-throughput sequencing count data. S. G. Walker is supported by the U. S. National Science Foundation through grant DMS-1506879. S. Favaro is supported by the European Research Council (ERC) through StG N-BNP 306406.

References

D. Aldous. Exchangeability and related topics. In *Ecole d'Ete de Probabilities de Saint-Flour XIII*, pages 1–198. Springer, 1983.

F. J. Anscombe. Sampling theory of the negative binomial and logarithmic series distributions. *Biometrika*, 37(3-4):358–382, 1950.

C. Antoniak. Mixtures of Dirichlet processes with applications to bayesian nonparametric problems. *Ann. Statist.*, (2):1152–1174, 1974.

J. G. Bethlehem, W. J. Keller, and J. Pannekoek. Disclosure control of microdata. *J. Amer. Statist. Assoc.*, 85(409):38–45, 1990.

D. Blackwell and J. MacQueen. Ferguson distributions via Pólya urn schemes. *Ann. Statist.*, 1(2): 353–355, 1973.

A. Brix. Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, 31(4):929–953, 1999.

A. Cerquetti. Generalized Chinese restaurant construction of exchangeable Gibbs partitions and related results. *arXiv:0805.3853*, 2008.

C. A Charalambides. *Combinatorial methods in discrete distributions*. Wiley, 2005.

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

L. J. Core, J. J. Waterfall, and J. T. Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–1848, 2008.

P. De Blasi, S. Favaro, A. Lijoi, R. H. Mena, I. Prunster, and M. Ruggiero. Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):212–229, 2015.

W. J. Ewens. *Theoretical population biology*, 3(1):87–112, 1972.

S. E. Fienberg and U. E. Makov. Confidentiality, uniqueness and disclosure limitation for categorical data. *Journal of Official Statistics*, 14(4):385–398, 1998.

R. A. Fisher, A. Steven Corbet, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology*, 12(1):42–58, 1943.

A. C. Frazee, B. Langmead, and J. T. Leek. Recount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, 12(1):449, 2011.

A. Gnedin and J. Pitman. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences*, 138(3):5674–5685, 2006.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

B. Greenberg and L. Voshell. The geographic component of disclosure risk for microdata. In *Statistical Research Division Report Series Census/SRD/RR-90/13, US Bureau of the Census*, 1990.

H. Ishwaran and L. F. James. Generalized weighted Chinese restaurant processes for species sampling mixture models. *Statistica Sinica*, pages 1211–1235, 2003.

N. L. Johnson, A. W. Kemp, and S. Kotz. *Univariate Discrete Distributions*. John Wiley & Sons, 2005.

J. F. C. Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

J. F. C. Kingman. Random partitions in population genetics. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 361, pages 1–20. The Royal Society, 1978a.

J. F. C. Kingman. The representation of partition structures. *Journal of the London Mathematical Society*, 2(2):374–380, 1978b.

J. F. C. Kingman. *Poisson Processes*. Oxford University Press, 1993.

R. M. Korwar and M. Hollander. Contributions to the theory of Dirichlet processes. *Ann. Probab.*, 1(4):705–711, 1973.

A. Lijoi and I. Prünster. Models beyond the Dirichlet process. In N. L. Hjort, C. Holmes, P. Müller, and S. G. Walker, editors, *Bayesian Nonparametrics*. Cambridge University Press, 2010.

Y. Liu, J. Zhou, and K. P. White. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, 30(3):301–304, 2014.

D. Manrique-Vallier and J. P. Reiter. Estimating identification disclosure risk using mixed membership models. *J. Amer. Statist. Assoc.*, 107(500):1385–1394, 2012.

M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5): 323–351, 2005.

M. Perman, J. Pitman, and M. Yor. Size-biased sampling of poisson point processes and excursions. *Probability Theory and Related Fields*, 92(1):21–39, 1992.

J. Pitman. *Combinatorial stochastic processes*. Lecture Notes in Mathematics. Springer-Verlag, 2006.

J. Pitman and M. Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Prob.*, 25(2):855–900, 1997.

M. H. Quenouille. A relation between the logarithmic, Poisson, and negative binomial series.

Biometrics, 5(2):162–164, 1949.

E. Regazzini, A. Lijoi, and I. Prünster. Distributional results for means of normalized random measures with independent increments. *Ann. Statist.*, 31(2):560–585, 2003.

D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, 2014.

C. J. Skinner and M. J. Elliot. A measure of disclosure risk for microdata. *J. R. Stat. Soc.: series B*, 64(4):855–867, 2002.

C. J. Skinner and N. Shlomo. Assessing identification risk in survey microdata using log-linear models. *J. Amer. Statist. Assoc.*, 103(483):989–1001, 2008.

M. Sultan, M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960, 2008.

Y. W. Teh. A Bayesian interpretation of interpolated kneser-ney. *NUS School of Computing Technical Report TRA2/06*, 2006.

Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

G. A. Watterson. Models for the logarithmic species abundance distributions. *Theoretical Population Biology*, 6(2):217–250, 1974.

F. Yang, T. Babak, J. Shendure, and C. M. Disteche. Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome research*, 20(5):614–622, 2010.

M. Zhou and L. Carin. Negative binomial process count and mixture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(2):307–320, 2015.

M. Zhou, O. H. M. Padilla, and J. G. Scott. Priors for random count matrices derived from a family of negative binomial processes. *To appear in J. Amer. Statist. Assoc.*, 2015.

G. K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley, 1949.

Table 1: Large n asymptotic regimes with respect to the parameter a.

a	Distinct types $l_{(n)}$	Distinct types $M_{i,n}$
$(0,1)$	$l_{(n)} \to 1 + \text{Poisson}\left(\frac{\gamma_0}{ap^a}\right)$	$M_{i,n} \to \text{Poisson}\left(\frac{\Gamma(i-a)\gamma_0 p^{-a}}{\Gamma(1-a)i!}\right)$
0	$\frac{l_{(n)}}{\log n} \to \gamma_0$	$M_{i,n} \to \text{Poisson}\left(\frac{\gamma_0}{i}\right)$
$-a \in \{1,2,\ldots\}$	$\frac{l_{(n)}}{n^{\frac{-a}{1-a}}} \to \frac{(\gamma_0 p^{-a})^{\frac{1}{1-a}}}{-a}$	$M_{i,n} \to \text{Poisson}\left(\frac{\Gamma(i-a)\gamma_0 p^{-a}}{\Gamma(1-a)i!}\right)$

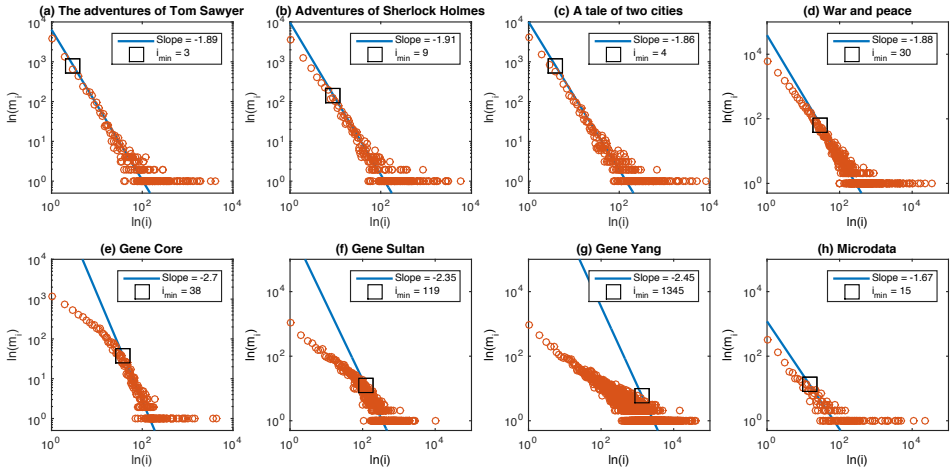

Figure 1: The log-log plots of the frequency of frequencies (FoF) vectors for (a) the words in "The Adventures of Tom Sawyer" by Mark Twain, (b) the words in "The Adventures of Sherlock Holmes" by Arthur Conan Doyle, (c) the words in "A Tale of Two Cities" by Charles Dickens, (d) the words in "War and Peace" by Leo Tolstoy and translated by Louise and Aylmer Maude, (e) the RNA sequences studied in Core et al. (2008), (f) the RNA sequences studied in Sultan et al. (2008), (g) the RNA sequences studied in Yang et al. (2010), and (h) the microdata provided in Table A.6 of Greenberg and Voshell (1990). For each subfigure, a least squares line with the slope fixed as $-\alpha$ is fitted to $\{[\ln i, \ln(m_i)]\}_{i:i\geq i_{\min}, m_i \geq 3}$, where $i_{\min}$ is a lower cutoff point and α is a scaling parameter estimated using the software provided for Clauset et al. (2009).

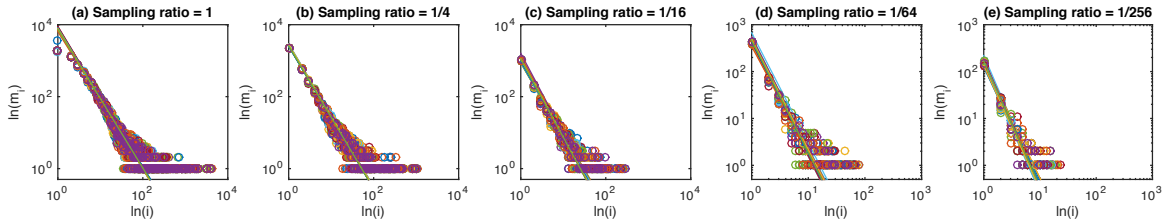

Figure 2: The log-log plots of the frequency of frequencies (FoF) vectors for the words in the novel "The Adventure of Tom Sawyer" by Mark Twain. Each subfigure consists of 20 FoF vectors displayed in different colors. (a) The 20 FoF vectors, with one curve coming from all the words and each of the other 19 curves coming from a sample of words taken with replacement from the novel, with a sampling ratio of 1; (b)-(e) The 20 FoF vectors, each of which comes from a sample of words taken without replacement from the novel, with the sampling ratios of 1/4, 1/16, 1/64, and 1/256, respectively. For each FoF vector, a straight line fitting the points $\{[\ln(i), \ln(m_i)]\}_{i:i\geq i_{\min}, m_i \geq 3}$ with slope $-\alpha$, is also plotted, where both the lower cutoff point $i_{\min}$ and scaling parameter α are estimated using the software provided for Clauset et al. (2009).

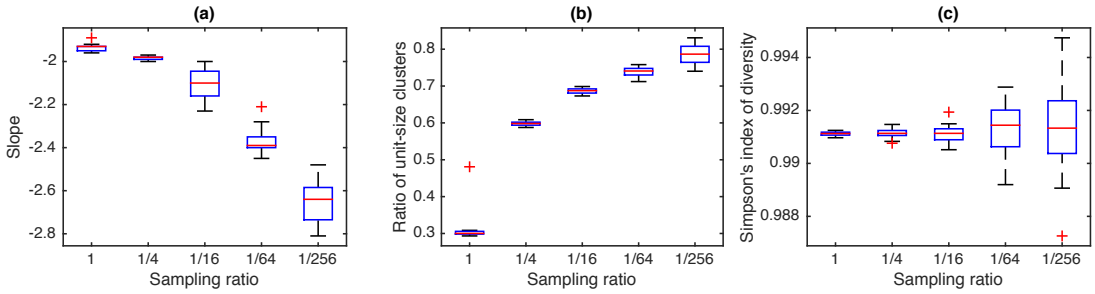

Figure 3: Box plots of (a) the slopes of the fitted lines and (b) the ratios of the clusters of size one for the FoF vectors in the log-log plots shown in Figure 2. For each sampling ratio, the box plot in each subfigure is based on the corresponding 20 FoF vectors used in Figure 2.

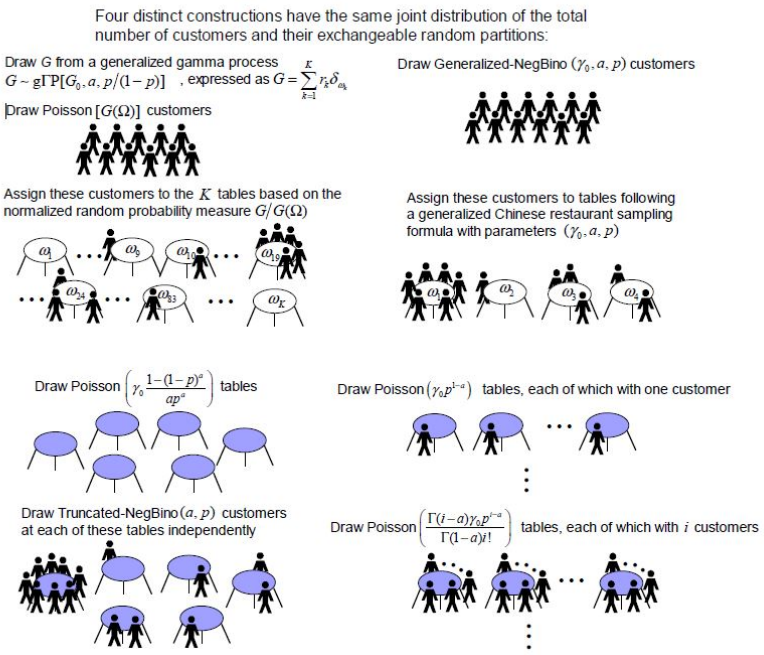

Figure 4: The cluster structure of the generalized negative binomial process can be either constructed by assigning Poisson$[G(\Omega)]$ number of customers to tables following a normalized generalized gamma process $G/G(\Omega)$, where $G \sim \mathrm{g\Gamma P}[G_0, a, p/(1-p)]$, or constructed by assigning $n \sim \mathrm{gNB}(\gamma_0, a, p)$ number of customers to tables following a generalized Chinese restaurant sampling formula $\boldsymbol{z} \sim \mathrm{gCRSF}(n, \gamma_0, a, p)$, where $\gamma_0 = G_0(\Omega)$. A equivalent cluster structure can be generated by first drawing Poisson$(\gamma_0 \frac{1-(1-p)^a}{ap^a})$ number of tables, and then drawing TNB(a, p) number of customers independently at each table. Another equivalent one can be generated by drawing Poisson$(\frac{\Gamma(i-a)\gamma_0 p^{i-a}}{\Gamma(1-a)i!})$ number of tables, each of which with i customers, for $i \in \{1, 2, \ldots\}$.

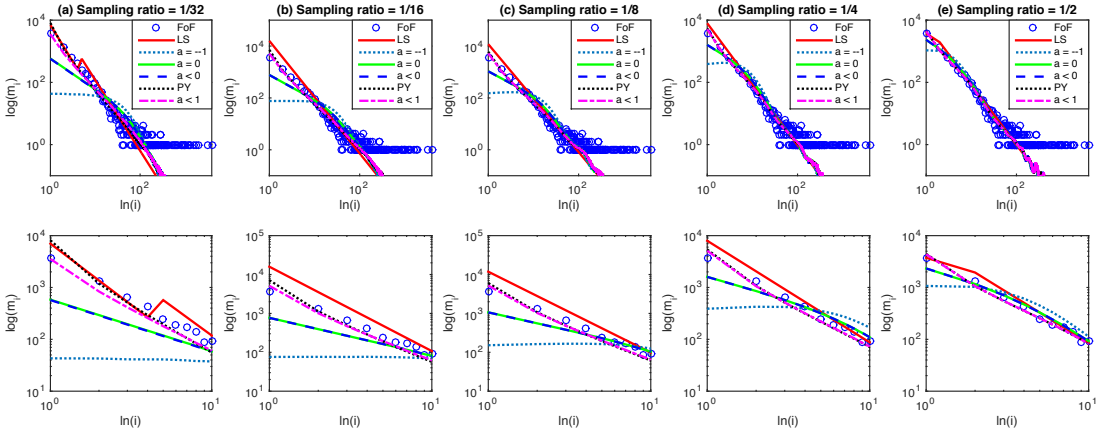

Figure 5: The posterior means of the population FoF vectors extrapolated from sample FoF vectors for "The Adventures of Tom Sawyer" by Mark Twain, using the least squares (LS) refitting procedure, the Chinese restaurant process, the Pitman-Yor (PY) process, and the generalized negative binomial process (gNBP), whose discount parameter is set as $a = -1$, $a = 0$, $a \in (-\infty, 0)$, or $a \in (-\infty, 1)$. Each sample is taken without replacement from the population with a sampling ratio of $1/32$, $1/16$, $1/8$, $1/4$, or $1/2$. The performance of the Chinese restaurant process is found to be almost identical to the gNBP with $a = 0$, and hence omitted for brevity.

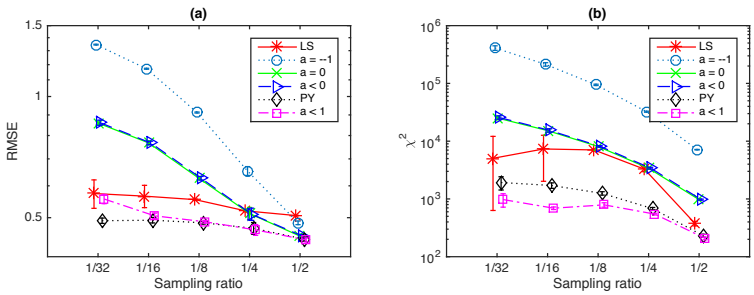

Figure 6: (a) RMSEs and (b) chi-squared (χ^2) test statistics for the extracted FoF vectors shown in Figure 5.

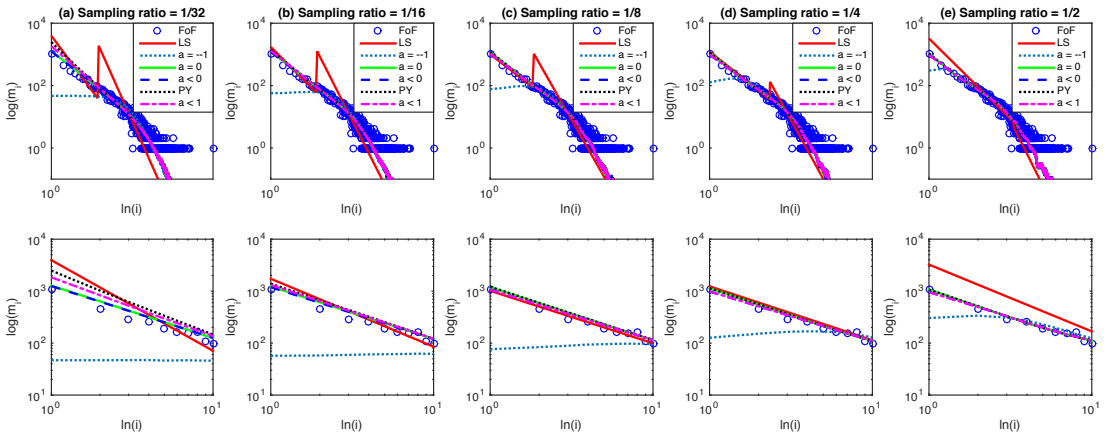

Figure 7: Analogous plots to Figure 5 for a RNA-seq data studied in Sultan et al. (2008).

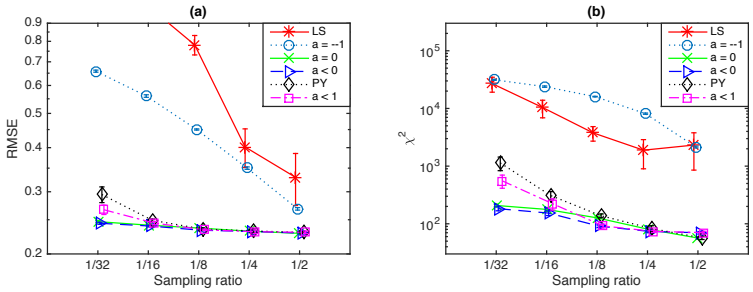

Figure 8: Analogous plots to Figure 6 for a RNA-seq data studied in Sultan et al. (2008).

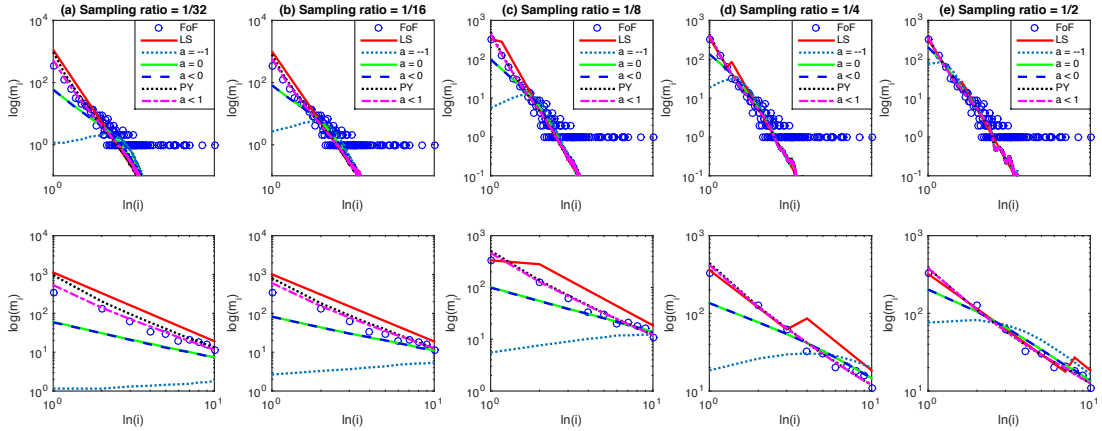

Figure 9: Analogous plots to Figure 5 for the microdata provided in Table A.6 of Greenberg and Voshell (1990).

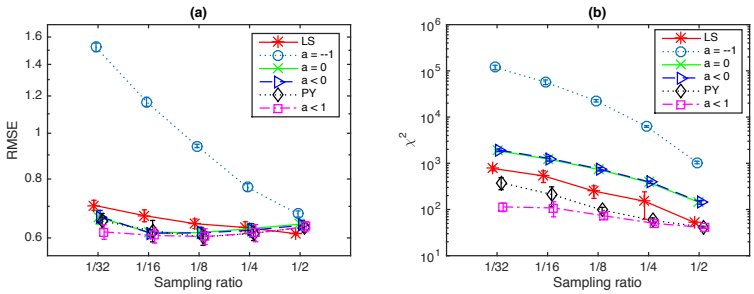

Figure 10: Analogous plots to Figure 6 for the microdata provided in Table A.6 of Greenberg and Voshell (1990).